Bayesian Analysis of Case-Control Studies

Bhramar Mukherjee, Samiran Sinha and Malay Ghosh
Department of Statistics, University of Florida, Gainesville FL 32611 USA
mukherjee@stat.ufl.edu, ssinha@stat.ufl.edu and ghoshm@stat.ufl.edu

Abstract

Case-control studies are dominant tools in epidemiologic research and consequently, there is a substantial volume of frequentist literature on different aspects of case-control studies. The Bayesian framework for case-control studies offer possibilities for flexible, hierarchical modeling that may be necessary in many contexts but as of yet, the Bayesian pathways remain relatively less explored. The present article reviews existing Bayesian work for analyzing case-control data, some recent advancements and possibilities for future research.

Keywords: Conditional likelihood, Dirichlet process, hierarchical model, likelihood inference, logistic regression, Mantel-Haenszel, Markov chain Monte Carlo, matching, missing data, measurement error, odds ratio, prospective study, retrospective study, semiparametric.

1 Introduction: The frequentist development

The contribution of statisticians to the development of case-control methodology is perhaps the most important contribution that they have made to public health and biomedicine. The central theme of a case-control study is to compare a group of subjects (cases) having the outcome (typically a disease) to a control group (not having the outcome or disease) with regard to one or more of the disease's potential risk factors. The method gained popularity in the 1920's for studying rare diseases, especially cancer, where following a healthy population or a cohort over time is impractical. It was believed in the early era of case-control studies that such a study did not provide relevant quantitative information about the disease rates. Cornfield (1951) demonstrated that the exposure odds ratio for cases versus controls equals the disease odds ratio for exposed versus unexposed, and the latter in turn approximates the ratio of the disease rates if the disease is rare. The following discussion will clarify the issue. Consider two dichotomous factors, D, denoting the disease status and E, representing the exposure status that characterize individuals in a population. The collected data is represented by a 2 × 2 table of the following form:

	D	$\bar{D}$
E	a	b
$\bar{E}$	c	d

One measure of association between D and E is the odds ratio (OR) defined as

$$OR = \frac{Pr(D|E)}{Pr(\bar{D}|E)} \cdot \frac{Pr(\bar{D}|\bar{E})}{Pr(D|\bar{E})}. \tag{1}$$

The disease odds ratio may be estimated from any of the three popular study designs namely, cohort, cross-sectional and case-control. It can be easily seen from (1) by using the Bayes theorem, namely, by using the fact that $Pr(E|D) = Pr(D|E)Pr(E)/Pr(D)$, that the exposure odds ratio and the disease odds ratio are the same. Another measure of association is the relative risk of the disease for different exposure values, defined as $RR = Pr(D|E)/Pr(D|\bar{E})$. When the disease is rare, both $Pr(\bar{D}|\bar{E})$ and $Pr(\bar{D}|E)$ are close to one and the exposure odds ratio approximates the relative risk of the disease (Cornfield (1951)).

The landmark paper by Mantel and Haenszel (1959) further clarified the relationship between a retrospective case-control study and a prospective cohort study. The paper considers a series of I 2×2 tables of the following pattern :

Disease Status	Exposed	Not Exposed	Total
Case	n_{11i}	n_{10i}	n_{1i}
Control	n_{01i}	n_{00i}	n_{0i}
Total	e_{1i}	e_{0i}	N_i

The Mantel-Haenszel (MH) estimator of the common odds ratio across the tables is

$$\hat{\theta}_{mh} = \frac{\sum_{i=1}^{I} n_{11i} n_{00i}/N_i}{\sum_{i=1}^{I} n_{01i} n_{10i}/N_i}. \tag{2}$$

and the test statistic to test $H_0 : \theta_1 = \cdots = \theta_I$ is $\sum_{i=1}^{I}\{n_{11i} - E(n_{11i}|\hat{\theta}_{mh})\}^2/Var(n_{11i}|\hat{\theta}_{mh})$ which follows an approximate χ^2 distribution with $I - 1$ degrees of freedom under the null hypothesis. The derivation of the variance of the MH estimator posed a challenge, and was addressed in several subsequent papers (see Breslow, 1996 for details).

The next era saw development of likelihood based inference methods for the odds ratio (Breslow and Day (1980)). Methods to evaluate the simultaneous effects of multiple quantitative risk factors on disease rates started appearing in 1960's.

In a case-control study the appropriate likelihood is the 'retrospective' likelihood of exposure given the disease status. Cornfield et al.(1961) noted that if the exposure distributions in case and control populations are normal with different means but a common covariance matrix, then the prospective probability of disease (D) given the exposure ($\boldsymbol{X}$) turns out to be the logistic response curve,

$$P(D = 1|\boldsymbol{X} = \boldsymbol{x}) = H(\alpha + \boldsymbol{\beta}^T \boldsymbol{x}) \tag{3}$$

where $H(u) = 1/\{1 + \exp(-u)\}$. Cox (1966), and Day and Kerridge (1967) confirmed that the maximum likelihood estimate of $\boldsymbol{\beta}$ obtained by the logistic regression model in (3) was efficient in a semiparametric sense. The only conceptual complication was in using a likelihood based on $P(D|\boldsymbol{X})$ whereas the case-control sampling design naturally leads to a likelihood containing terms of the form $P(\boldsymbol{X}|D)$. Anderson (1972) and Prentice and Pyke (1979) investigated this problem and showed that the ordinary logistic regression coefficients actually yielded the desired estimates of the relative risk parameters. They also established that the estimating equations based on the prospective logistic likelihood is unbiased and the asymptotic standard error of the resulting estimators of the relative risk parameters are identical with those obtained from the retrospective model. Carroll, Wang and Wang (1995) extended this prospective formulation to the situation of missing data and measurement error in the exposure variable. They showed that the estimator of $\boldsymbol{\beta}$ obtained through prospective formulation and ignoring the retrospective nature of the case control data is consistent and asymptotically normally distributed with prospective standard errors that are at worst conservative.

In a case-control set-up, matching is often used for selecting "comparable" controls to eliminate bias due to confounding factors. Statistical techniques for analyzing matched case-control data were first developed in Breslow et al.(1978). In the simplest setting, the data consist of s matched sets and there are M_i controls matched with a case in each matched set or stratum. We denote the i-th matched set by S_i, $i = 1, \cdots, s$. As before, one assumes a prospective stratified logistic disease incidence model, namely,

$$P(D = 1|\boldsymbol{z}, S_i) = H\{\alpha_i + \boldsymbol{\beta}^T(\boldsymbol{z} - \boldsymbol{z}_0)\} \tag{4}$$

where α_i's are stratum-specific intercept terms. The stratum parameters α_i are eliminated by conditioning on the unordered set of exposures for the cases and controls in each stratum. This is equivalent to conditioning on the number of cases in each stratum which is a complete sufficient statistic for the nuisance parameters α_i. The generated conditional likelihood, is free of the nuisance parameters and yields the optimum estimating function (Godambe, 1976) for estimating $\boldsymbol{\beta}$. Assuming, without loss of generality, the first subject in each stratum is a case and rest of the subjects are controls, the derived conditional likelihood is

$$L_c(\boldsymbol{\beta}) \;=\; \prod_{i=1}^{s} \frac{\exp(\boldsymbol{\beta}^T \boldsymbol{z}_{i1})}{\sum_{j=1}^{M_i+1} \exp(\boldsymbol{\beta}^T \boldsymbol{z}_{ij})}. \qquad (5)$$

The above method is known as conditional logistic regression(**CLR**).

The usual prospective likelihood for unmatched data and conditional logistic regression for matched studies produce biased estimates of the relative risk parameters in the presence of measurement error or misclassification in the exposure variable. Armstrong, Whittemore and Howe (1989) proposed a correction to the estimates obtained through the classical methods using normal discriminant analysis with normal differential measurement error (to be made precise in Section 2). Carroll, Gail and Lubin (1993) proposed a pseudo-likelihood method to handle this situation, where they assume a differential measurement error density with a finite set of parameters, and empirical distribution functions are used in place of needed distributions. Recently, McShane et al. (2001) proposed a conditional score method for estimating bias-corrected estimates of log odds ratio parameters in matched case control studies.

The classical methods for analyzing unmatched and matched studies suffer from loss of efficiency when the exposure variable is partially missing. Lipsitz, Parzen and Ewell (1998) proposed a pseudo-likelihood method to handle missing exposure variable. Rathouz, Satten and Carroll (2002) developed a more efficient semiparametric method of estimation in presence of missing exposure in matched case control studies. Satten and Kupper (1993), Paik and Sacco (2000) and Satten and Carroll (2000) addressed the missingness problem from a full likelihood approach assuming a distribution of the exposure variable in the control population.

The above account of the frequentist development of case-control analysis is nowhere near complete and is beyond the scope of the current article. Here we only attempt to review the Bayesian contributions in this field. In spite of the enormously rich literature in the frequentist domain, Bayesian modeling for case-control studies did not start until the late 1980's. Bayesian analysis of case-control data seem particularly appealing with the rapid development of Markov chain Monte Carlo techniques. The possibilities include random effects, measurement error, missingness, flexibility to incorporate hierarchical structure and prior information in modeling the relative risk. In fact, epidemiology is indeed a science where accumulated knowledge from similar prior studies should be utilized in the analysis of a new one. In the subsequent three sections we attempt to review the available Bayesian literature on modeling data arising from case-control studies. Sections 2 and 3 focus on models for unmatched case-control data. In Section 2, we consider models for a single binary exposure. Section 3 contains a review of more involved modeling of continuous exposure with measurement error and categorical covariates. Section 4 is devoted to methods available for the analysis of matched case-control studies. Section 5 focuses on some recent work on equivalence of prospective and retrospective studies in a Bayesian framework and discusses a new Bayesian interpretation to the method of conditional likelihood. Section 6 presents concluding remarks and possible directions for future research.

2 Early Bayesian work on a single binary exposure

Altham (1971) is possibly the first Bayesian paper which considered several 2×2 contingency tables with a common odds ratio, and performed a Bayesian test of association based on this common odds ratio. This paper is not specifically targeted towards a case-control design. Later, Zelen and Parker (1986), Nurminen and Mutanen (1987), Marshall (1988) and Ashby et al. (1993) all considered identical Bayesian formulations of a case-control model with a single binary exposure X. The model can be described as follows:

Let ϕ and γ be the probabilities of exposure in control and case populations respectively. The retrospective likelihood is

$$l(\phi, \gamma) \propto \phi^{n_{01}}(1-\phi)^{n_{00}}\gamma^{n_{11}}(1-\gamma)^{n_{10}}. \tag{6}$$

Where n_{01} and n_{00} are the number of exposed and unexposed observations in control population, whereas n_{11} and n_{10} denotes the same for case population.

Independent conjugate prior distributions for ϕ and γ are assumed to be $Beta(u_1, u_2)$ and $Beta(v_1, v_2)$ respectively. After reparametrization one obtains the posterior distribution of the log odds ratio parameter, namely, $\beta = log\{\gamma(1-\phi)/\phi(1-\gamma)\}$ as

$$p(\beta|n_{11}, n_{10}, n_{01}, n_{00}) \propto \exp\{(n_{11}+v_1)\beta\} \int_0^1 \frac{\phi^{n_{11}+n_{01}+v_1+u_2-1}(1-\phi)^{n_{10}+n_{00}+v_2+u_1-1}}{\{1-\phi+\phi\exp(\beta)\}^{n_{11}+n_{10}+v_1+v_2}} d\phi. \quad (7)$$

The above posterior density of β does not exist in closed form but may be evaluated by numerical integration.

Since interest often lies in the hypothesis $\beta = 0$, Zelen and Parker (1986) recommended calculating the ratio of the two posterior probabilities $p(\beta)/p(0)$ at selected deviates β. When β is set at the posterior mode, a large value of this ratio will indicate concentration of the posterior away from 0 and one would infer disease-exposure association. However the "critical value" suggested for this ratio to detect presence of association is completely arbitrary. Zelen and Parker also provided a normal approximation to the posterior distribution of β to avoid numerical computation, and discussed the problem of choosing a prior distribution based on prior data available on exposure values. Using their prior elicitation method and normal approximation to the posterior, they analyzed data showing association between exposure in utero to Diesthylstibestrol (a drug which came in 1940's to prevent pregnancy complications) and adenocarcinoma of vagina in young women. The ratio of the posterior density at the posterior mode and at 0 turn out to be overwhelmingly large at 302, suggesting strong evidence of association between disease and exposure. This paper also raised the possibility of performing a case-control analysis without the presence of control sample if adequate exposure information is available from prior studies.

Nurminen and Mutanen (1987) considered a more general parameterization based on a general comparative parameter such as the risk ratio or the risk difference, and not just the odds ratio. They provided a complicated exact formula for the cumulative density function of this general comparative parameter. The formula can be related to Fisher's exact test for comparing two proportions in sampling theory. The Bayesian point estimates are considered as posterior median and mode, whereas inference is based on the highest posterior density interval for the comparative parameter of interest.

Marshall (1988) provided a closed form expression for the moments of the posterior distribution of the odds ratio. The paper mentions that an approximation to the exact posterior density of the odds ratio parameter can be obtained by power series expansion of the hypergeometric functions involved in the expression for the density, but acknowledge the problem of slow convergence in adopting this method. Marshall then used Lindley's (1964) result on the approximate normality of $\log(OR)$ which works very well over a wide range of situations. In the absence of any exposure information, he recommended using uniform independent priors for the parameters. The paper suggests that a perception about the value of the odds ratio should guide the choice of prior parameters rather than attempting to exploit the exposure proportions as suggested in Zelen and Parker. Inference again is based on posterior credible intervals. Marshall (1988) reanalyzed the Zelen-Parker data on association of vaginal cancer and Diesthylstibestrol (DES). With the same prior as Zelen and Parker, the posterior credible interval for the odds ratio turned out to be (7.8, 388). Since this was a dataset with very few observations, the author chose a more conservative set of priors which also suggested association between vaginal cancer and DES . The paper concludes that the exact magnitude of the risk may not be quantified realistically based on such a small dataset.

Ashby et al.(1993) analyzed a case-control study to assess risk of leukemia following chemotherapy for Hodgkin's disease from a Bayesian perspective and used it as a source of prior information for a second study. They also proposed modeling the odds ratio directly through a normal normal model besides the usual beta-binomial modeling. The paper emphasized practical relevance of Bayesian perspective from a epidemiological standpoint. Epidemiology progresses by an accumulation of evidence about the existence and magnitude of risk factors; so a Bayesian approach offers a natural framework for integrating and updating the knowledge available at each stage.

In cancers for which survival is good, there is a growing interest in the long-term effects of treatment. Hodgkin's disease, which is cancer of the lymph nodes, has a 5 year survival of about 80% and survivors have excess risks for leukemias and lymphomas. Although these risks are very small, compared to the benefits of the treatment, there is interest in studying these effects. The paper considers three such studies: an initial cohort study (Kaldor *et al.*, 1987), an international case-control study conducted in 1990 and a case-control study finished in 1993 as a follow-up to the 1990 study with the same protocol as the earlier study. The data for the 1990 case-control study is analyzed with no prior information, slightly informative prior and then using the 1987 cohort study as prior evidence. Then a fraction of data collected in a follow-up case-control study is analyzed with the initial 1990 study as prior information. With small amount of data, prior information plays a vital role. Interesting numerical differences may be noted for different choices of the prior. The paper makes a very clear point that before embarking on a study of a possible etiological risk for any disease, the investigators may consider existing evidence across a range of sources for arriving at a more efficient strategy for design and analysis of the data following Bayesian principles.

3 Models with continuous and categorical exposure

Müller and Roeder (1997) introduced a novel aspect to Bayesian treatment of case-control studies by considering continuous exposure with measurement error. This paper is a significant advance in Bayesian analysis of case-control data as it is the first attempt to address non-trivial, unorthodox modeling scenarios beyond the earlier work on binary exposure as discussed in Section 2.

Müller and Roeder used the validation group, for which error-free high quality measurements are available to form a model that indirectly links the error-prone measurement to the disease outcome through its relation to the error-free measurement. The set-up of their paper is as follows:

Let D be a binary response observed together with covariates X and Z, where Z may be vector-valued. The retrospective case-control sample is chosen by selecting n_1 cases with

$D = 1$ and n_0 controls with $D = 0$. Errors in variables occur when for a subset of the data, (referred as reduced data or $\mathcal{R}$ from now on) a proxy W is measured instead of the true covariate X. For a typical validation study, for the remaining subjects (referred to as complete data or $\mathcal{C}$ from now on) one has observations on both the gold standard X and the proxy W.

The approach proposed in their paper is based on a nonparametric model for the retrospective likelihood of the covariates and the imprecisely measured exposure. For reduced data, this nonparametric distribution models the joint distribution of $(Z$,$W)$ and the missing covariate X_R. For complete data, the same distribution models the joint distribution of the observed covariates (X_C, Z, W). The nonparametric distribution is a class of flexible mixture distributions, obtained by using mixture of normal models with a Dirichlet Process prior on the mixing measure (Escobar and West, 1995). The prospective disease model relating disease to exposure is taken as a logistic distribution characterized by a vector of log odds-ratio parameters $\boldsymbol{\beta}$. Also, if $\boldsymbol{\theta}$ denotes the parameters describing the marginal distributions of the exposure and covariates, under certain conditional independence type assumptions and non-differential measurement error, i.e. $p(D|X, Z, W, \boldsymbol{\beta}) = p(D|X, Z, \boldsymbol{\beta})$, the retrospective likelihood is shown to be compatible with the prospective model.

With a prior $p(\boldsymbol{\beta}, \boldsymbol{\theta})$, one obtains the joint posterior as

$$p(\boldsymbol{\beta}, \boldsymbol{\theta}|X_C, W, D, Z) \propto p(\boldsymbol{\beta}, \boldsymbol{\theta}) \prod_{i \in \mathcal{C}} p(X_i, Z_i, W_i|D_i, \boldsymbol{\beta}, \boldsymbol{\theta}) \prod_{i \in \mathcal{R}} \{ \int p(X_i, Z_i, W_i|D_i, \boldsymbol{\beta}, \boldsymbol{\theta}) dX_i \}$$

By augmenting the parameter with the latent vector one can write the joint posterior of parameters and missing exposure X_R as,

$$p(\boldsymbol{\beta}, \boldsymbol{\theta}, X_R|X_C, W, D, Z) \propto p(\boldsymbol{\beta}, \boldsymbol{\theta}) \prod_{i=1}^{n} p(X_i, Z_i, W_i|D_i, \boldsymbol{\beta}, \boldsymbol{\theta}).$$

Under the additional assumption of non-differential measurement error,

$$p(\boldsymbol{\beta}, \boldsymbol{\theta}, X_R|X_C, W, D, Z) \propto p(\boldsymbol{\beta}, \boldsymbol{\theta}) \prod_{i=1}^{n} p(X_i, Z_i, W_i|\boldsymbol{\theta}) p(D_i|X_i, Z_i, \boldsymbol{\beta})/p(D_i|\boldsymbol{\beta}, \boldsymbol{\theta})$$

where $p(D_i|\boldsymbol{\beta},\boldsymbol{\theta}) = \int p(D_i, X_i, Z_i, W_i, \boldsymbol{\beta}) dP(X_i, Z_i, W_i|\boldsymbol{\theta})$. Let $\boldsymbol{\theta} = (\theta_1, \theta_2, .., \theta_n)$. Then one assumes the following mixture model with a multivariate normal kernel ϕ_{θ_i}. Specifically.

$$\begin{aligned} p(X_i, Z_i, W_i|\theta_i) &= \phi_{\theta_i}(X_i, Z_i, W_i) \\ \theta_i &\sim G \\ G &\sim DP(\alpha G_0) \end{aligned} \tag{8}$$

Where θ_i =(μ_i, Σ_i), the parameters of the trivariate normal kernel and $DP(\alpha G_0)$ denotes a Dirichlet process with a base measure G_0 and concentration parameter α. The hierarchy is completed by assuming a normal/Wishart hyperprior on the parameters of the base measure $G_0 = N(\mu_0, \Sigma_0)$ and a Gamma prior on the concentration parameter α. A non-informative prior is assumed on $\boldsymbol{\beta}$. A Markov chain Monte Carlo numerical integration scheme is designed for estimating the parameters. The computation scheme becomes either very expensive or inaccurate when the dimension of the (X, Z, W)-space increase as one has to evaluate $p(D_i|\boldsymbol{\beta},\boldsymbol{\theta})$ by numerical integration over the (X, Z, W)-space. This paper breaks many new grounds in Bayesian treatment of case-control studies including consideration of continuous covariates, measurement error and flexible nonparametric modeling for the exposure distribution leading to robust, interpretable results quantifying the effect of exposure. This paper brings out the tremendous possibility of using modern Bayesian computational techniques to offer solutions to complex data scenarios in case-control studies.

Müller and Roeder analyzed a dataset extracted from Lipids Research Clinic's prospective study to estimate the risk of coronary heart disease given levels of LDL cholesterol and triglycerides (TRG). An individual is classified as a case if he has experienced a previous heart attack, an abnormal exercise electrocardiogram or a history of angina pectoris.

Since direct measurement of LDL is expensive, total cholesterol (TC) is often used as a proxy. In their set-up, onset of coronary disease, log(TC), log(LDL) and log(TRG) play the roles of D, W, X and Z respectively. Based on their analysis with and without the reduced observations, it is clear that the reduced observations are highly informative and TC is an excellent proxy for LDL. The posterior distribution for the log-odds ratio parameter β_1 corresponding to X (LDL) has a much smaller variance when the reduced data is used.

Müller and Roeder pointed out that categorical covariates required a different treatment as assuming a multivariate normal kernel for the Dirichlet Process implicitly assumes continuous exposures. Seaman and Richardson (2001) extended the binary exposure model of Zelen-Parker to any number of categorical exposures simply by replacing the binomial likelihood in (6) by a multinomial likelihood, and then adopting a MCMC strategy to estimate the log odds ratio for the disease in each category with respect to a baseline category. The set of multinomial probabilities corresponding to exposure categories in case and control populations are assumed to have a discrete Dirichlet prior. The prior on the log odds ratios could be chosen to be any distribution.

Seaman and Richardson also adapted the Müller-Roeder approach to categorical covariates by simply modeling the multinomial probabilities corresponding to exposure categories in the entire source population as opposed to looking at case and control populations separately as in the generalized Zelen-Parker binary model described above. A non-informative Dirichlet (0,0,..,0) prior is assumed for the exposure probabilities. Seaman and Richardson establish the equivalence of the generalized binary and adapted Müller-Roeder in some limiting cases with noninformative Dirichlet (0,0,..,0) prior on ϕ.

Continuous covariates may be treated in the Seaman and Richardson framework by discretizing them into groups and little information is lost if discretization is sufficiently fine. This opens up the possibility of treating continuous and categorical covariates simultaneously which is not obvious in the Müller-Roeder approach.

Seaman and Richardson as well as Müller and Roeder assumed a prospective logistic likelihood, assumed a flexible prior for the exposure distribution, and derived the implied retrospective likelihood. In contrast, Müller et al. (1999) specified a retrospective likelihood, and then derived the implied prospective likelihood. Müller et al. (1999) specified both $p(X|D = 0, \theta)$ and $p(X|D = 1, \theta)$ as mixtures of multivariate normal distributions with

M components. First we consider the simple normal case and begin with the notation, $p(\boldsymbol{X}|D, \boldsymbol{\theta}_D) = N(\boldsymbol{\mu}_D, \boldsymbol{\Sigma}_D)$, where $\boldsymbol{\theta}_D = (\boldsymbol{\mu}_D, \boldsymbol{\Sigma}_D)$ for $D = 0$ and 1; so θ_0 and θ_1 are parameters of the exposure distributions in the control and case populations. Then we denote the mixtures of multivariate normal distributions as $p(\boldsymbol{X}|D, \boldsymbol{\theta}) = \sum_{m=0}^{M} \pi_{Dm} N(\boldsymbol{\mu}_{Dm}, \boldsymbol{\Sigma}_{Dm})$ where $\boldsymbol{\theta}_D = (\pi_{D0}, \boldsymbol{\mu}_{D0}, \boldsymbol{\Sigma}_{D0}, \cdots, \pi_{DM}, \boldsymbol{\mu}_{DM}, \boldsymbol{\Sigma}_{DM})$ for $D = 0$ and 1. These authors assumed $\gamma = P(D = 1)$, a constant prevalence of the disease. With this retrospective formulation, the following can be shown:

Result. (i) In the simple normal case, i.e., when $\boldsymbol{\mu}_{Dm} = \boldsymbol{\mu}_D$ and $\boldsymbol{\Sigma}_{Dm} = \boldsymbol{\Sigma}_D$ for D=0,1, the prospective log odds have a quadratic form $\eta_p(\boldsymbol{X}) = \boldsymbol{X}'\boldsymbol{A}\boldsymbol{X} + \boldsymbol{X}'b + c$, with $2\boldsymbol{A} = \boldsymbol{\Sigma}_0^{-1} - \boldsymbol{\Sigma}_1^{-1}$, $b = \boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0$ and $2c = \boldsymbol{\mu}_0'\Sigma_0^{-1}\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1'\Sigma_1^{-1}\boldsymbol{\mu}_1 + 2log\{\gamma/(1-\gamma)\} + log(|\Sigma_0|/|\Sigma_1|)$.

(ii) In the simple normal model, if in addition $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_0$, then the prospective model is the logistic with linear predictor $\eta_p(\boldsymbol{X}) = \boldsymbol{X}'b + c$.

(iii) In the mixture of normals model, the prospective log odds are of the form $\eta_p(\boldsymbol{X}) = log\{\sum_{m=0}^{M} \pi_{1m}^* \exp(\boldsymbol{X}'\boldsymbol{A}_{1m}\boldsymbol{X} + \boldsymbol{X}'\boldsymbol{b}_{1m} + c_{1m})\} - log\{\sum_{m=0}^{M} \pi_{0m}^* \exp(\boldsymbol{X}'\boldsymbol{A}_{0m}\boldsymbol{X} + \boldsymbol{X}'\boldsymbol{b}_{0m} + c_{0m})\} + log\{\gamma/(1-\gamma)\}$, where $\pi_{Dm}^* = \pi_{Dm}|\boldsymbol{\Sigma}_{Dm}|^{-\frac{1}{2}}$, $\boldsymbol{A}_{Dm} = -\frac{1}{2}\boldsymbol{\Sigma}_{Dm}^{-1}$, $\boldsymbol{b}_{Dm} = \boldsymbol{\Sigma}_{Dm}^{-1}\boldsymbol{\mu}_{Dm}$ and $c_{Dm} = -\frac{1}{2}\boldsymbol{\mu}_{Dm}^t\boldsymbol{\Sigma}_{Dm}^{-1}\boldsymbol{\mu}_{Dm}$ for $D = 0$, 1.

Result (ii) reiterates the classical result of Cornfield (1961) about discriminating the exposure distributions in case and control populations using the prospective logistic model when the exposures have a multivariate normal distribution with different means but common covariance matrix in the case and control populations. Result (iii) is a generalization which states that with the mixture of M normals retrospective model, the prospective odds ratio turn out to be a weighted average of M quadratic regressions in the log odds which becomes a weighted average of M linear regressions when the case and control distributions have the same dispersion parameter. Their example shows that non-linear modeling of the odds ratio as described in (iii) may be needed in some real situations. The choice of M, the number of components in the mixture model is guided by two competing principles, easier estimation and flexibility of modeling. In their example, Müller et al. use M=3.

Müller et al. (1999) also addressed the problem of handling categorical and quantitative exposure simultaneously. Conditional on the quantitative covariates, a multivariate probit model is assumed for the binary covariates. They mentioned that this model could be generalized for general categorical covariates as well, but did not specify exactly how this might be done. They developed MCMC scheme for implementation of this model and for inference and prediction. The computational difficulty of evaluating the normalizing constant in using the prospective model as seen in Müller-Roeder is circumvented here through direct modeling of the retrospective likelihood. Two other attractive features of this paper are 1) to incorporate inter-center variability while combining data from eight tumor registries and 2) provide a natural framework of combining data coming from case-control and cohort studies. For the latter, the likelihood is based on both prospectively and retrospectively collected individuals. The risk factors for a retrospectively ascertained individual enter the likelihood conditionally on whether the person is a case or a control. For a prospectively ascertained individual, the disease outcome and individual risk factors are modeled jointly in terms of $p(\boldsymbol{X}, D|\boldsymbol{\theta}, \gamma)$ where $\gamma = p(D = 1)$ is the disease prevalence. Let P be the set of indices corresponding to all prospectively ascertained patients. Then the combined likelihood can be written as:

$$L(\boldsymbol{\theta}, \gamma) = \prod_{D_i=1} p(\boldsymbol{X}_i|D = 1, \boldsymbol{\theta}) \prod_{D_i=0} p(\boldsymbol{X}_i|D = 0, \boldsymbol{\theta}) \prod_{i \in P} \gamma^{D_i}(1-\gamma)^{1-D_i}$$

The example using the Cancer and Steroid Hormone (CASH) study indicates that the two commonly adopted assumptions regarding the linearity of the log odds and homogeneity of log odds ratio across studies need not always be true. Models based on non-linear and non-homogeneous odds ratios may lead to a more complete account of uncertainties than the usual logistic regression model.

Recently, Gustafson et al.(2002) considered another Bayesian approach to handle exposure variables that are measured imprecisely. They considered approximation of the imprecise exposure distribution by a discrete distribution supported on a suitably chosen grid. In the absence of measurement error, the support is chosen as the set of observed values of the exposure, a device that resembles the Bayesian Bootstrap (Rubin, 1981). In the presence of measurement error, these authors proposed strategies for selection of these grid points depending on the validation design. A Dirichlet(1,...,1) prior was assigned to the probability vector corresponding to the set of grid points. The posterior distributions of the parameters of interest were derived directly from the retrospective likelihood of data and design, and a MCMC computing scheme was used for carrying out posterior inference. Interestingly, in the process of devising the MCMC algorithm, they showed that the posterior distribution as derived from a prospective likelihood can be used as an approximation to the posterior distribution directly derived from the retrospective likelihood. They also proposed an importance sampling scheme based directly on the posterior derived from retrospective data. It is observed that if the importance weights exhibit little variation, then the inference based on the prospective model suffices. In the subsequent data analysis, very small differences are noted between prospective and retrospective models for moderate sample sizes and the authors continue to work with the prospective model which is more amenable to the MCMC analysis. This paper is the first one to shed some light on the interplay between the equivalence of prospective and retrospective studies in a Bayesian perspective which we continue to discuss in Section 5. The equivalence established in Gustafson et al. (2002) is an approximate equivalence whereas Seaman and Richardson (2004) established exact equivalence results (see Section 5).

4 Analysis of matched case-control studies

All the articles discussed so far in the Bayesian domain considered unmatched case-control studies. One requirement for a case-control study is that the selection of cases and controls is independent of their exposure status. In the simplest form of study designs, cases and controls are identified and information on exposure and relevant confounder variables (age, sex, race) are determined on the sampled subjects. The subsequent analysis tries to adjust for the confounding effects by including the additional terms in the model or by doing a stratified Mantel-Haenszel analysis. The design could potentially be somewhat inefficient if large number of controls fall within confounder-defined strata which contain only a few cases. In such a situation the control information is effectively wasted. Matched case-control designs try to circumvent this problem by keeping the ratio of cases to controls constant within each sub-group defined by the relevant confounders, called the *matching variables*. Frequency matching corresponds to maintaining constant case-control ratios across broad strata whereas individual matching selects a set of matched controls for each sampled case. As discussed in Section 1, frequentist inference in the matched set-up proceeds via conditional likelihood . However, Bayesian attempts to analyze a matched case-control study have started appearing only very recently.

Diggle, Morris and Wakefield (2000) present the first Bayesian analysis for situations when cases are individually matched to the controls and as a result, nuisance parameters are introduced to represent the separate effect of matching in each matched set. Diggle et al. (2000) consider matched data when exposure of primary interest is defined by the spatial location of an individual relative to a point or line source of pollution.

The basic model of Diggle et al. (2000) for an unmatched set-up is as follows: Let the locations of individuals in the population at risk follow an nonhomogeneous Poisson process with intensity function $g(x)$ and the probability that an individual at location x becomes a case is $p^*(x)$. Then the odds of disease amongst sampled individuals are:

$$r(x) = (a/b)p^*(x)/(1 - p^*(x)),$$

where a and b are the sampled proportions of cases and controls. Typically, $r(x) \equiv r(x, \theta)$ involves some unknown parameter θ, and is modeled as

$$r(x, \theta) = \rho h(x, \theta)$$

where $h(x, \theta)$ involves interpretable parameters of interest θ. The extension of this model to J individually matched case-control pairs in a study region is made in the following way. Suppose, the J locations for cases and controls are denoted by x_{0j} and x_{1j} respectively, j=1,..,J. The probability of disease for an individual at location x in stratum j is modeled as:

$$p_j(x,\theta) = \tfrac{r_j(x,\theta)}{1+r_j(x,\theta)} = \tfrac{\rho_j h(x,\theta)}{1+\rho_j h(x,\theta)},$$

where ρ_j are varying baseline odds between matched pairs and are considered as nuisance parameters. Conditioning on the unordered set of exposures within each matched set, one has the conditional probability of disease of an individual at distance x in stratum j as,

$$p_c(x_{j0},\theta) = \tfrac{h(x_{j0},\theta)}{h(x_{j0},\theta)+h(x_{j1},\theta)}$$

In case of $1:M$ matching in each stratum and in presence of q additional covariates $z_k(x_{ji})$ measured for the i-th individual in j-th stratum at location x_{ji}, $i = 1,..M; j = 1,..J; k = 1,..q$, the full fledged conditional probability can be modeled in the form:

$$p_c(x_{j0},\theta,\phi) = \tfrac{h(x_{j0},\theta)\exp(\sum_{k=1}^{q} z_k(x_{j0})\phi_k)}{\sum_{i=0}^{M} h(x_{ji},\theta)\exp(\sum_{k=1}^{q} z_k(x_{j0})\phi_k)},$$

with the conditional log likelihood being of the form:

$$L(\theta,\phi) = \sum_{j=1}^{J} \text{log} p_c(x_{j0},\theta,\phi).$$

Diggle et. al. (2000) considered the likelihood as well as Bayesian approach for estimation of parameters and inference in this model with specified form of $h(x,\theta)$. Often in studying environmental pollution, there is a putative point source at location x^* and interest is on how the risk surface changes in relation to x^*. Let $d = \|x - x^*\|$ be the distance of location x from the source. In such instances they propose h of the form

$$h(x) = 1 + \alpha\exp(-(d/\beta)^2), \text{ for fixed } x^*.$$

This model has a natural interpretation with α being the proportional increase in disease odds at the source, while β measures the rate of decay with increasing distance from the source in units of distance. Standard likelihood inference for such models based on the likelihood ratio statistic may not be reliable as the likelihood involved is highly non-regular.

The Bayesian paradigm provides an alternative when likelihood surface is highly irregular. In the paper independent normal priors are assumed on the covariate related regression parameters ϕ and independent uniform priors are chosen for α and β. Estimation is carried out by adopting a MCMC technique and inference is conducted by computing the Bayes factor for testing the null hypothesis $h(x) = 1$ which amounts to saying that the risk of the disease does not depend on the distance from the putative source.

Diggle *et al.* analyzed two sets of data, collected to study the relationship between proximity of residence to main roads and increased risk of respiratory disease. The first dataset consisted of 125 asthma cases, and the second 124 chronic obstructive airways disease (COAD). Cases were clear-cut diagnoses attending the casualty deprament at the Royal London Hospitals in 1991-92. Controls were selected from non-chest related admissions. Each case was individually matched to a single control using sex, race and consultant team. The analysis was based on the distance odds model as described above. For the Asthma data, no association could be detected between asthma and distance of residence to main roads. The COAD data exhibited a reasonable evidence of elevated risk close to main roads, but the functional form of the variation in risk with distance could not be estimated with precision. The paper indicates introducing a stochastic spatial component in the model which could be an interesting direction for further research. This paper indicates using conditional likelihood as a basis for Bayesian inference and what assumptions are implicitly made regarding the choice of priors when such an analysis is carried out. This issue is investigated in detail by Rice (2004) as discussed in Section 5.

Ghosh and Chen (2002) developed general Bayesian inferential techniques for matched case-control problems in the presence of one or more binary exposure variables. The model considered was more general than that of Zelen and Parker (1986). Also, unlike Diggle, Morris, and Wakefield (2000), the analysis was based on an unconditional likelihood rather than a conditional likelihood after elimination of nuisance parameters. Any methodology based on the conditional likelihood is applicable only for the logit link. The general Bayesian methodology based on the full likelihood as proposed by Ghosh and Chen worked beyond the logit link. Their procedure included not only the probit and the complementary log links but also some new symmetric as well as skewed links. The propriety of posteriors was proved under a very general class of priors which need not always be proper. Also, some robust priors such as multivariate t-priors were used. The Bayesian procedure was implemented via Markov chain Monte Carlo.

To illustrate their general methodology, Ghosh and Chen considered the L.A. Cancer data as given in Breslow and Day (1980), and studied the effect of gall-bladder disease on the risk of endometrial cancer as a single exposure variable, and both gall-bladder disease and hypertension as multiple exposure variables. In the case of one case and one control, it was demonstrated that the Bayesian procedure could yield conclusions different from the conditional likelihood regarding association between the exposures and the disease. This is because, while the conditional likelihood ignores all concordant pairs, the Bayesian procedure based on the full likelihood utilizes these pairs as well.

Rice (2003) considered a matched case-control analysis where a binary exposure is potentially misclassified. His method can be interpreted in a Bayesian framework with prior information incorporated on odds ratios, misclassification rates and other model parameters.

Recently Sinha et al. (2004, 2005a,b) have proposed a unified Bayesian framework for matched case-control studies with missing exposure and a semiparametric alternative for modeling varying stratum effects on the exposure distribution. They considered s, $1:M$ matched strata with a disease indicator D, a vector of completely observed covariate $\boldsymbol{Z}$ and an exposure variable X with possible missing values. The missingness is assumed to be at random (MAR in the sense of Little and Rubin (1987)). They assume a prospective model for the disease status as,

$$P(D_{ij}=1|X_{ij},\boldsymbol{Z}_{ij},S)=H(\beta_0(S)+\boldsymbol{\beta}_1^T\boldsymbol{Z}_{ij}+\beta_2 X_{ij}),$$

where ij refers the j^{th} individual in the i^{th} stratum, $j=1,\cdots,M+1$ and $i=1,\cdots,s$. The distribution of the exposure variable X in a control population is assumed to be a member of general exponential family, namely

$$p(X_{ij}|\boldsymbol{Z}_{ij}, S_i, D_{ij} = 0) = \exp[\xi_{ij}\{\theta_{ij}X_{ij} - b(\theta_{ij})\} + c(\xi_{ij}, X)] \tag{9}$$

where the canonical parameter θ_{ij} is modeled in terms of the completely observed covariate $\boldsymbol{Z}_{ij}$ and stratum specific varying intercept terms, namely $\theta_{ij} = \gamma_{0i} + \boldsymbol{\gamma}^T\boldsymbol{Z}_{ij}$. They propose two important lemmas which are useful to write down the likelihood.

Lemma 1. $\frac{\mathrm{pr}(D_{ij}=1|\boldsymbol{Z}_{ij},S_i)}{\mathrm{pr}(D_{ij}=0|\boldsymbol{Z}_{ij},S_i)} = \exp\left[\beta_0(S_i) + \boldsymbol{\beta}_1^T\boldsymbol{Z}_{ij} + \xi_{ij}\{b(\theta^*_{ij}) - b(\theta_{ij})\}\right]$, where $\theta^*_{ij} = \theta_{ij} + \xi_{ij}^{-1}\beta_2$.

Lemma 2. Based on the above two model assumptions, one can derive the distribution of the exposure variable in the case population as:

$$p(X_{ij}|D_{ij} = 1, \boldsymbol{Z}_{ij}, S_i) = \exp[\xi_{ij}\{\theta^*_{ij}X_{ij} - b(\theta^*_{ij})\} + c(X_{ij}, \xi_{ij})]. \tag{10}$$

Using the above two lemmas and assuming data missing at random, the prospective joint conditional likelihood of the disease status and the exposure variable given the completely observed covariate, the stratum effect and the conditional event that there is exactly one case in each stratum can be shown to be proportional to,

$$\begin{aligned}
& \prod_{i=1}^{s} P(D_i = 1, D_{i2} = \cdots D_{iM+1} = 0, \{X_{ij}, \Delta_{ij}\}_{j=1}^{M+1}|S_i, \{\boldsymbol{Z}_{ij}\}_{j=1}^{M+1}, \sum_{j=1}^{M+1} D_{ij} = 1) \\
\propto\ & \prod_{i=1}^{s}\Bigg\{P(D_{i1} = 1, D_{i2} = \cdots = D_{iM+1} = 0|\{\boldsymbol{Z}_{ij}\}_{j=1}^{M+1}\sum_{j=1}^{M+1} D_{ij} = 1, S_i) \\
& \times p^{\Delta_{i1}}(X_{i1}|\boldsymbol{Z}_{i1}, D_{i1} = 1, S_i) \times \prod_{j=2}^{M+1} p^{\Delta_{ij}}(X_{ij}|\boldsymbol{Z}_{ij}, D_{ij} = 0, S_i)\Bigg\} \\
=\ & \prod_{i=1}^{s}\Bigg\{\frac{P(D_{i1} = 1|\boldsymbol{Z}_{i1}, S_i)/P(D_{i1} = 0|\boldsymbol{Z}_{i1}, S_i)}{\sum_{j=1}^{M+1} P(D_{ij} = 1|\boldsymbol{Z}_{ij}, S_i)/P(D_{ij} = 0|\boldsymbol{Z}_{ij}, S_i)} \\
& \times p^{\Delta_{i1}}(X_{i1}|\boldsymbol{Z}_{i1}, D_{i1} = 1, S_i) \times \prod_{j=2}^{M+1} p^{\Delta_{ij}}(X_{ij}|\boldsymbol{Z}_{ij}, D_{ij} = 0, S_i)\Bigg\},
\end{aligned}$$

In the above expression it is assumed that the first observation in each stratum is the one coming from case population, Δ_{ij} is the missing value indicator and takes value 0 if $\boldsymbol{X}_{ij}$ is missing and 1 otherwise.

The parameters were estimated in a Bayesian framework by using a non-parametric Dirichlet Process prior on the stratum specific effects γ_{0i} in the distribution of the exposure variable and parametric priors for all other parameters. Since the number of these stratum specific parameters grow with the sample size, estimating these effects from a conditional MLE perspective may lead to the Neyman-Scott phenomenon. The novel feature of the Bayesian semiparametric method is that it can capture unmeasured stratum heterogeneity in the distribution of the exposure variable in a robust manner. The proposed method has been extended to the situation when one has multiple disease states (see Sinha, Mukherjee and Ghosh, 2004). The method is appealing as a unified framework as one can not only handle the missingness in the exposure variable but also possible measurement error and misclassification in the exposure variable. The method may also be extended to take into account the possible association schemes that may exist in a mixed set of continuous and categorical multiple exposures with partial missingness (Sinha, Mukherjee and Ghosh, 2005b). Their examples and simulation results indicate that in presence of missingness, if association among exposures truely exist, one gains efficiency in estimating the relative risk parameters by modeling the exposure association instead of ignoring it.

In the following subsections we present some of the examples presented in Sinha et al. (2004, 2005a) of Bayesian analysis of matched case control study.

4.1 A continuous exposure: The equine epidemiology example

The first example is equine epidemiology example earlier analyzed by Kim, Cohen and Carroll (2002). The data consist of 498 strata with 1 : 1 matching in each stratum. Participating veterinarians were asked to provide data monthly for one horse treated for colic and one horse that received emergency treatment for any condition other than colic, between March 1, 1997, and February 28, 1998. A case of colic was defined as the first horse treated during a given month for signs of intra-abdominal pain. A control horse was defined as the next horse that received emergency treatment for any condition other than colic. Age was considered as a single exposure variable (X) measured on a continuous scale with one binary covariate (Z) indicating whether the horse experienced recent diet changes or not. For scaling purposes the authors linearly transformed age so that X was on the interval $[0, 1]$. The proposed method for the general exponential family, as described in the previous section, applies here with continuous exposure X. Sinha *et al.* (2005a) assumed that conditional on $D_{ij} = 0$, i.e., for the control population the exposure variable age has a normal distribution of the following form :

$$[X_{ij}|D_{ij} = 0, Z_{ij}, S_i] \sim \text{Normal}(\gamma_{0i} + \gamma_1 Z_{ij}, \sigma^2), \tag{11}$$

where γ_{0i} are stratum effects on the exposure distribution and γ_1 is the regression parameter for regressing X_{ij} on the measured covariates Z_{ij}. They also assumed the logistic regression model for disease status. From Lemma 2 it follows the exposure distribution among the cases is Normal with mean $\gamma_{0i} + \gamma_1 Z_{ij} + \sigma^2\beta_2$ and variance σ^2. The conditional likelihood can now be derived using these facts. Suitable priors are chosen on the parameters. For estimation of the parameters, a Markov chain Monte Carlo numerical integration scheme was used. For simulating observations from the posterior distribution of the γ_{0i}, the methods proposed in West, Müller and Escober (1994), Escober and West (1995) and Neal (2000) was used. For generating random numbers from the conditional distribution of other parameters, the authors followed a componentwise Metropolis-Hastings scheme.

For each of the examples as presented in sections 4.1-3 and in Sinha et al. (2005a), three analyses were conducted. One is the proposed Bayesian semiparametric analysis (BSP). The other is a parametric Bayesian (PB) version of the work of Satten and Carroll (2000) by considering a constant stratum effect $\gamma_{0i} \equiv \gamma_0$ and then assuming a normal prior on this common stratum effect parameter γ_0. They emphasize here and in the tables that the parametric Bayesian method is simply the Bayesian version of the method of Satten and Carroll (2000). The authors also present the frequentist alternative for analyzing this matched data through a standard conditional logistic regression (CLR) analysis (Breslow and Day, 1980).

They ran the three analyses mentioned above on the equine dataset with all 498 strata and reran the analyses where they randomly deleted 40% of the age observations. The results are given in Table 1.

Briefly, with no missing data the results of the methods are roughly in accordance with one another. The slightly small standard error estimates for the Bayesian methods may be a reflection of the fact that unlike CLR, they are full likelihood based semiparametric and parametric methods. With missing data, of course, one can see that the Bayesian methods yield much smaller standard errors than CLR, reflecting the fact that their methods use all the data, and not just the pairs with no missing data. There is little difference between the semiparametric analysis and the parametric Bayes version of the method of Satten and Carroll (2000), a phenomenon that can be explained as follows. The average variance among the γ_{0i}'s is very small (about 0.0054), suggesting that the stratum effects are almost constant across strata. Thus the parametric model with constant stratum effect essentially holds for this example, and hence it is not surprising that the two methods give similar results.

4.2 A binary exposure: endometrial cancer study

In the Los Angeles 1:4 matched case-control study on endometrial cancer as discussed in Breslow and Day (1980), the variable obesity may be considered as a binary exposure variable (X) for contracting endometrial cancer. This variable had about 16% missing observations. Sinha *et al.* (2005a) considered the presence of gall bladder disease in the subject as our completely observed dichotomous covariate Z.

In such a case with dichotomous exposure variable, the exposure distribution for control population is naturally assumed to have a logistic form: $\mathrm{pr}(X_{ij} = 1|Z_{ij}) = H(\gamma_{0i} + \gamma_1 Z_{ij})$, where $H(\cdot)$ is the logistic distribution function. From Lemma 2 the exposure distribution in the case population is again of the logistic form with: $\mathrm{pr}(X_{ij} = 1|Z_{ij}) = H(\gamma_{0i} + \gamma_1 Z_{ij} + \beta_2)$.

In the LA study there were 63 strata. The original data contained missing observations on the exposure variable obesity. They modeled the missingness in the exposure variable in both the parametric Bayes (the Bayesian version of the Satten and Carroll method with fixed stratum effects) and the semiparametric Bayes analysis. They also conducted classical conditional logistic analysis. Table 2 presents the results. Although there are certainly some interesting numerical differences, in the main the results are reasonably comparable.

The average variance among the γ_{0i}'s over the last 5000 MCMC runs is approximately 4.5. This is fairly large, and if there were major covariate effects one would expect the BSP and the parametric Bayes version of the method of Satten and Carroll (2000) to be different. However, in this example the analyses suggest that $\beta_2 = 0$ is plausible, so that D and X are essentially independent and whether the γ_{0i}'s are constant or not does not cause a model bias. This is an example where there is stratum variability and natural missingness in the data.

4.3 Another binary exposure: low birthweight study

In Example 4.1 on equine data, there was no appreciable stratum effect variability though the exposure had a significant effect on disease probability. Example 4.2 on the other hand has natural missingness, appreciable stratum variability but the exposure variable obesity did not have significant effect on disease probability. The example considered in this section illustrates a case when there is significant stratum variability and modeling it causes some interesting differences in inference about the exposure effect when compared to a constant stratum effect model.

This example involves a matched case-control dataset coming from a low birth weight study conducted by the Baystate Medical Center in Springfield, Massachusetts. The dataset is discussed in Hosmer and Lemeshow (2000, Section 1.6.2) and is used as an illustrative example of analyzing a matched case-control study in Chapter 7 of their book. Low birth weight, defined as birth weight less than 2500 grams, is a cause of concern for a newborn as infant mortality and birth defect rates are very high for low birth weight babies. The data was matched according to the age of the mother. A woman's behavior during pregnancy (smoking habits, diet, prenatal care) can greatly alter the chances of carrying the baby to terms. The goal of the study was to determine whether these variables were “risk factors” in the clinical population served by Baystate Medical Center. The matched data contain 29 strata and each stratum has one case (low birthweight baby) and 3 controls (normal birthweight baby). One can possibly think of many different models for explaining the disease in terms of the possible covariates recorded in the data set. The authors considered smoking status of mother as a single exposure variable. Two other covariates, a binary variable denoting presence of uterine irritability (UI) in mother and weight of the mother at last menstrual period (LWT) are also included in the model.

In this example the average stratum variability is about 3.82 justifying the need for a varying stratum effect model even in the absence of any missingness. Consequently, one can note in Table 3 that in the full data case, the significance of smoking is more pronounced with a varying stratum effect model when compared to the CLR and the parametric Bayes version of Satten and Carroll which assumes a constant stratum effect. The latter two models only exhibit marginal significance of the exposure. The association between smoking and low birth weight has been clinically well-established (Walsh, 1994) though the exact mechanism of decreased birth-weight is not yet known. With introduction of missingness, the BSP estimate for the log odds ratio in Table 4 for smoking remains very close to the full data counterpart whereas the estimates from the constant stratum effect models change appreciably.

4.4 Example of a matched case-control study with multiple disease states

For this example Sinha *et al.*, (2004) again considered the low birth weight data discussed in section 4.3. They divided the cases, namely, the low birth-weight babies into two categories,*very* low (weighing less than 2000 gms) and low (weighing between 2000 to 2500 gms) and tried to assess the impact of smoking habits of mother on the chance of falling in the two low birth-weight categories relative to the baseline category (normal birthweight, weighing more than 2500 grams). Presence of uterine irritability in mother and mother's weight at last menstruation period were considered as relevant covariates. It was noted that smoking mothers had a higher relative risk of having a low birth weight child when compared to a non-smoking mother. However, the risk of having a *very* low birth weight child did not depend on smoking significantly.

For the analysis the conditional probabilities of the disease variable given the covariate, exposure and the stratum are given by,

$$P(D_{ij} = k | S_i, \mathbf{Z}_{ij}, X_{ij}) = \frac{\exp\{\beta_{0k}(S_i) + \boldsymbol{\beta}_{1k}^T \mathbf{Z}_{ij} + \beta_{2k} X_{ij}\}}{1 + \sum_{r=1}^{K} \exp\{\beta_{0r}(S_i) + \boldsymbol{\beta}_{1r}^T \mathbf{Z}_{ij} + \beta_{2r} X_{ij}\}} \text{ for } k = 1, \cdots, K. \quad (12)$$

and binary variable smoking status is assumed to follow a Bernoulli distribution. In their example $K = 2$. Three analyses were performed, namely Bayesian Semiparametric (BSP), Parametric Bayes (PB) and the third one is iid parametric (PBV) where they assumed all the stratum effects are different and each of the n stratum effect parameters γ_{0i}, $i = 1, \cdots, n$ are coming from n independent normal distributions.

Table 5 contains the posterior means, posterior standard deviations and 95% HPD credible intervals for the parameters of interest under the proposed Bayesian Semiparametric method (BSP) and the parametric Bayes (PB the PBV) methods as discussed before. To illustrate the methods in presence of missingness, the authors deleted 40% of exposure values completely at random and reran the three analyses. The results are presented in Table 6. The full data analysis indicates that smoking of mother is a significant risk factor for low birth-weight category (category 1) and is not very significant in the *very* low birth-weight category (category 2). UI on the other hand shows an opposite association, showing significance in category 2 and almost no significance in category 1. LWT does not seem to be a significant covariate in any of the categories. The BSP and the PBV methods are in closer agreement whereas the PB estimates show some numerical differences. Obviously, without the finer classification into two weight categories, the fact that smoking is not so significant for category 2 and UI is appreciably significant for category 2 cannot be concluded from looking at the overall analysis.

For the analysis with 40% missing observations on smoking one notices that the estimates corresponding to smoking in the BSP method comes closer to their full data counterparts even though the inferences are same in all three methods. As one might expect, with 40% missingness, the parameter estimates for smoking lose precision and the effect of smoking now appears to be insignificant in both categories 1 and 2. Inferences on the other two covariates remain essentially unchanged when compared to full data inferences.

5 Some equivalence results in case-control studies

5.1 Equivalence of retrospective and prospective analysis

Prentice and Pyke (1979) showed that if the disease risk is modeled by logistic regression function (3) and the subjects are selected into the study irrespective of their exposure value, prospective and retrospective analysis of the case-control data yield the same estimate of the association parameter $\boldsymbol{\beta}$. Moreover, the asymptotic standard error of the estimate are the same in both the methods. However, the intercept parameter of the prospective model of the disease risk is not identifiable in a retrospective likelihood unless we know the disease prevalence in the population. Generally, the prospective model involves fewer parameters, and hence is easy to implement. Prentice and Pyke (1979) provided the theoretical validity for prospective analysis of data collected retrospectively.

Roeder, Carroll, and Lindsay (1996) proved a similar result when exposure variables are measured with error. They showed that the profile likelihood function of the association parameter obtained from a retrospective likelihood is the same as the one obtained through the joint distribution of the disease variable, true exposure variable, and its error prone surrogate variable. Carroll, Wang and Wang (1995) extended Prentice-Pyke type results to situations with missingness and measurement error in the exposure variable.

Recently, Seaman and Richardson (2004) proved a Bayesian analogue of the Prentice-Pyke equivalence result. They considered a discrete exposure vector $\boldsymbol{X}$ with J support points $\{z_1, \ldots, z_J\}$. Let n_{0j} and n_{1j} be the number of cases and controls having the exposure value $\boldsymbol{X} = \boldsymbol{z}_j$, $j = 1, \cdots, J$. Now if $\text{pr}(\boldsymbol{X} = \boldsymbol{z}_j | D = 0) = \theta_j / \sum_{j=1}^{J} \theta_j$, and odds of disease associated with $\boldsymbol{X} = \boldsymbol{x}$ is $\exp(\boldsymbol{\beta}^T \boldsymbol{x})$, then the natural retrospective likelihood for the case-control data is

$$L_{MR} = \prod_{d=0}^{1} \prod_{j=1}^{J} \Big\{ \frac{\theta_j \exp(d\boldsymbol{\beta}^T \boldsymbol{z}_j)}{\sum_{k=1}^{J} \theta_k \exp(d\boldsymbol{\beta}^T \boldsymbol{z}_k)} \Big\}^{n_{dj}}. \tag{13}$$

If one assumes that the data came from a cohort study, then the natural prospective likelihood is

$$L_{MP} = \prod_{j=1}^{J} \prod_{d=0}^{1} \Big\{ \frac{\alpha^d \exp(d\boldsymbol{\beta}^T \boldsymbol{z}_j)}{\sum_{k=0}^{1} \alpha^k \exp(d\boldsymbol{\beta}^T \boldsymbol{z}_j)} \Big\}^{n_{dj}}, \tag{14}$$

where α is the baseline odds of disease when exposure $\boldsymbol{X} = \mathbf{0}$. The authors proved the following two results.

Theorem 1. The profile likelihood of $\boldsymbol{\beta}$ obtained by maximizing L_{MR} with respect to $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_J)$ is the same as the profile likelihood of $\boldsymbol{\beta}$ obtained by maximizing L_{MP} with respect to α.

The next theorem proves the Prentice-Pyke type equivalence result in Bayesian context for a specific class of priors.

Theorem. Suppose that random variables n_{dj} $(d = 0, 1; j = 1, \cdots, J)$ are independently distributed as $n_{dj} \sim Po(\lambda_{dj})$, where $\log\lambda_{dj} = d\log\alpha + \log\theta_j + d\boldsymbol{\beta}^T \boldsymbol{z}_j$. Assume independent improper priors, $p(\alpha) \propto \alpha^{-1}$ and $p(\theta_j) \propto \theta_j^{a_j - 1}$, for α and $\boldsymbol{\theta}$, and a prior, $p(\boldsymbol{\beta})$ for $\boldsymbol{\beta}$ that is independent of α and $\boldsymbol{\theta}$. Let $n_{+j} = n_{0j} + n_{1j}$, and $n_{d+} = \sum_{j=1}^{J} n_{dj}$. Also, let $\boldsymbol{n}^T = (n_{01}, \cdots, n_{0J}, \cdots, n_{11}, \cdots, n_{1J})$.

(i) The joint posterior density of $(\omega, \boldsymbol{\beta})$, where $\omega = \log\alpha$, is

$$p(\omega, \boldsymbol{\beta} | \boldsymbol{n}) \propto p(\boldsymbol{\beta}) \prod_{j=1}^{J} \frac{\{\exp(\omega + \boldsymbol{\beta}^T \boldsymbol{z}_j)\}^{n_{1j}}}{\{1 + exp(\omega + \boldsymbol{\beta}^T \boldsymbol{z}_j)\}^{n_{+j} + a_j}}. \tag{15}$$

(ii) The posterior density of $(\boldsymbol{\pi}, \boldsymbol{\beta})$, where $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_J)$ and $\pi_j = \theta_j / \sum_{k=1}^{J} \theta_k$ is

$$p(\boldsymbol{\pi}, \boldsymbol{\beta} | \boldsymbol{n}) \propto p(\boldsymbol{\beta}) \prod_{j=1}^{J} \pi_j^{a_j - 1} \prod_{d=0}^{1} \Big[\frac{\prod_{j=1}^{J} \{\pi_j \exp(d\boldsymbol{\beta}^T \boldsymbol{z}_j)\}^{n_{dj}}}{\{\sum_{j=1}^{J} \pi_j \exp(d\boldsymbol{\beta}^T \boldsymbol{z}_j)\}^{n_{d+}}} \Big]. \tag{16}$$

The proportionality is upto a constant.

Corollary: The marginal posterior densities of $\boldsymbol{\beta}$ obtained from the joint densities in (15) and (16) are the same.

The above corollary states that one may fit either the prospective model or the retrospective model and obtain the same posterior marginal distribution for the parameter of interest $\boldsymbol{\beta}$. One can note that the prospective model contains only one nuisance parameter α, whereas the retrospective model contains J nuisance parameters. Consequently, the prospective model is easier to fit.

The equivalence result of Seaman and Richardson (2004) is a significant contribution to the Bayesian literature on case-control studies. However, the results are limited in the sense that they only apply to the Dirichlet prior on the exposure distribution in the control population. The equivalence results for any other type of prior still an open question. Another limitation is that the exposure variable is assumed to be discrete. For continuous exposure, the authors recommend categorization which is not an ideal solution. The extension of the results to individually matched case-control studies under missingness and/or measurement error in exposure values may provide deeper insight into the Bayesian equivalence of prospective and retrospective model formulation.

5.2 Equivalence between conditional and marginal likelihood for analyzing matched case-control data

The disease risk model (4) for matched case-control study involves a set of nuisance parameters α_i's which capture the stratification effect on the disease risk. A natural way of analyzing matched case-control data is the conditional likelihood approach, where one considers conditional likelihood of the data conditioned on an approximately ancillary statistic for the parameter of interest (see e.g. Godambe (1976)). In a matched case-control study we condition on the complete sufficient statistic for the nuisance parameters α_i, namely, the number of cases in each matched set. Alternatively, one can work with a marginal likelihood obtained by integrating the nuisance parameters over a mixing distribution, say G. Rice (2004) showed that the full conditional likelihood can exactly be recovered via marginal likelihood approach by integrating the nuisance parameter with respect to a particular mixing distribution G. The author derived the necessary and sufficient conditions on the class of distributions G for the two approaches to agree. The conditions invoke certain invariance properties of the distribution G and such invariant distributions are shown to exist under certain mild natural condition on the odds ratios. In the light of this agreement, in a Bayesian framework, the posterior distribution of the disease-exposure association parameter, $\boldsymbol{\beta}$, of a matched case-control study using an invariant mixture distribution as a prior on the nuisance parameter and a flat prior on the association parameter, is proportional to the conditional likelihood. This provides a validation of using informative priors together with conditional likelihood as done in Diggle, Morris and Wakefield (2000).

6 Conclusion

In this review article we have focused on Bayesian modeling and analysis of case-control studies. The Bayesian paradigm offers a great deal of flexibility to accommodate unusual, unorthodox data situations and incorporating prior information on risk related parameters but comes with certain computational challenges. The popularity and use of these methods is highly dependent on developing user friendly softwares for implementing the analysis.

There are other issues besides analysis, like that of Bayesian variable selection (Raftery and Richardson, 1996), sample size determination (De Santis et al., 2001) in case-control studies which are not discussed in this article but are interesting in their own right.

Case-control studies bear enough promise for further research, both methodological and applied. One potential area of research is to extend the methodology when longitudinal data are available for both cases and controls. Hierarchical Bayesian modeling, because of its great flexibility, should prove to be a very useful tool for the analysis of such data. A second important area of research is to adapt the method for the analysis of survival data. Some work in this regard has been initiated in a frequentist framework by Prentice and Breslow (1978), but Bayesian inference for such problems seems to be largely unexplored.

For example the methods could potentially be adapted to bivariate survival models in family based study design (see Oakes, 1986, 1989 and Shih and Chatterjee, 2002). Finally, genetic case-control study is a new emerging area of research where one of the objectives is to study the association between a candidate gene and a disease. In many such instances, the population under study is assumed to be homogeneous with respect to allele frequencies, but comprises subpopulations that have different allele frequencies for the candidate gene (see Satten et. al., 2001). If these subpopulations also have different disease risks, then an association between the candidate gene and disease may be incorrectly estimated without properly accounting for population structures. Accounting for population stratification may pose some interesting statistical challenges in case-control studies of disease-gene or disease-gene-environment association.

Bayesian approach could also be useful to analyze data coming from other designs used in an epidemiologic context (Khoury, Beaty and Cohen (1993)). For example, another frequently used design is a nested case-control design (Wacholder, 1996) where a case control study is nested within a cohort study. This type of study is useful to reduce the cost and labor involved in collecting the data on all individuals in the cohort as well as to reduce computational burden associated with time-dependent explanatory variable. Unlike the case-control design this design allows us to estimate the absolute risk of the disease by knowing the crude disease rate from the cohort study. Some other study designs along this line are the case cohort design (Prentice, 1986), proband design (Gail, Pee and Carroll (2001)) and case-only design (Armstrong, 2003). In a recent article, Cheng and Chen (2005) propose Bayesian analysis of case-control studies of genetic association, specifically for assessing gene-environment interaction via case-only design under independence of genetic and environmental factors. They use informative power prior (Ibrahim and Chen (2000)) which in their analysis is taken to be the retrospective likelihood based on historic data raised to a suitable power.

It is important to note that case-control designs are choice-based sampling designs in which the population is stratified on the values of the response variable itself. Among others, this was noticed in Scott and Wild (1986) who compared in such cases a maximum likelihood based approach with some other ad hoc methods of estimation. Breslow and Cain (1988) considered in such cases a two-phase sampling design. In later years, there is a long series of publications on inference for such designs, but any Bayesian approach to these problems is still lacking, and will be a worthwhile future undertaking.

Full Equine Data					
Method		β_1	β_2	$10 \times \gamma_1$	$10 \times \sigma^2$
Bayes Semiparametric					
	Mean	2.16	2.18	0.18	0.23
	s.e.	0.33	0.39	0.14	0.01
	Lower HPD	1.57	1.36	-0.10	0.21
	Upper HPD	2.80	2.88	0.40	0.25
Bayes Parametric					
	Mean	2.14	2.10	0.22	0.26
	s.e.	0.32	0.41	0.17	0.01
	Lower HPD	1.60	1.32	-0.20	0.24
	Upper HPD	2.87	2.88	0.60	0.28
CLR					
	Mean	2.13	2.05		
	s.e.	0.32	0.47		
	Lower CL	1.50	1.13		
	Upper CL	2.76	2.97		

Equine Data With 40% Missing Data					
Method		β_1	β_2	$10 \times \gamma_1$	$10 \times \sigma^2$
Bayes Semiparametric					
	Mean	2.16	2.30	0.14	0.23
	s.e.	0.32	0.48	0.24	0.01
	Lower HPD	1.70	1.30	-0.25	0.19
	Upper HPD	2.80	3.42	0.60	0.25
Bayes Parametric					
	Mean	2.13	1.90	0.17	0.28
	s.e.	0.32	0.47	0.20	0.02
	Lower HPD	1.56	0.92	-0.20	0.24
	Upper HPD	2.85	2.85	0.67	0.31
CLR					
	Mean	2.81	2.79		
	s.e.	0.59	0.77		
	Lower CL	1.65	1.47		
	Upper CL	3.96	4.49		

Table 1: Results of the Equine Data example. Here “Mean” is the posterior mean, “s.e.” is the posterior standard deviation, “Lower HPD” and “Upper HPD” are the lower and upper end of the HPD region respectively,“Lower CL” and “Upper CL” are the lower and upper end of the confidence limit respectively, The parametric Bayesian method is the Bayesian version of the method of Satten and Carroll (2000).

		Gall Bladder	Obesity	
Method		β_1	β_2	γ_1
Bayes Semiparametric				
	Mean	1.29	0.70	-0.15
	s.e.	0.39	0.41	0.49
	Lower HPD	0.54	-0.10	-0.99
	Upper HPD	2.18	1.49	0.99
Bayes Parametric				
	Mean	1.19	0.49	0.42
	s.e.	0.39	0.36	0.62
	Lower HPD	0.58	-0.15	-0.70
	Upper HPD	2.12	1.25	1.78
CLR				
	Mean	1.28	0.44	
	s.e.	0.39	0.38	
	Lower CL	0.52	-0.30	
	Upper CL	2.04	1.19	

Table 2: Results of the Endometrial Cancer Data example. Here “Mean” is the posterior mean, “s.e.” is the posterior standard deviation, “Lower HPD” and “Upper HPD” are the lower and upper end of the HPD region respectively,“Lower CL” and “Upper CL” are the lower and upper end of the confidence limit respectively, The parametric Bayesian method is the Bayesian version of the method of Satten and Carroll (2000).

	BSP			PBC			CLR	
Parameter	Mean	Sd	HPD region	Mean	Sd	HPD region	Estimate	S.E
SMOKE	1.21	0.56	(0.15,2.31)	0.96	0.46	(-0.01,1.85)	0.86	0.45
LWT	-1.17	1.26	(-3.39,1.78)	-1.34	1.25	(-3.75,1.15)	-1.13	1.36
UI	0.91	0.49	(0.03,1.98)	0.79	0.52	(-0.31,1.86)	0.85	0.51

Table 3: Full data analysis of low birth weight data. BSP stands for Bayesian semiparametric method whereas PBC stands for parametric Bayes method assuming constant stratum effects. CLR stands for usual conditional logistic regression for analyzing matched data.

	BSP			PBC			CLR	
Parameter	Mean	Sd	HPD region	Mean	Sd	HPD region	Estimate	S.E
SMOKE	1.19	0.82	(-0.45,2.85)	0.65	0.59	(-0.47,1.98)	0.65	0.53
LWT	-1.12	1.27	(-3.51,1.46)	-1.22	1.24	(-3.70,1.21)	-1.06	1.62
UI	0.96	0.55	(-0.01,2.20)	0.86	0.53	(-0.30,1.91)	0.88	0.60

Table 4: Analysis of low birth weight data with 30% missingness in the SMOKE variable. BSP stands for Bayesian semiparametric method whereas PBC stands for parametric Bayes method assuming constant stratum effects. CLR stands for usual conditional logistic regression for analyzing matched data.

		BSP			PB			PBV		
Logit	Parameter	Mean	Sd	HPD region	Mean	Sd	HPD region	Mean	Sd	HPD
1	SMOKE	1.42	0.60	(0.33 ,2.72)	1.26	0.56	(0.25 ,2.50)	1.48	0.65	(0.26,2.08)
	LWT	-0.86	1.39	(-3.78 ,1.81)	-1.03	1.35	(-3.58 ,1.86)	-0.73	1.36	(-3.40,2.01)
	UI	0.15	0.67	(-1.27 ,1.46)	0.10	0.67	(-1.19 ,1.52)	0.18	0.67	(-1.14,1.52)
Logit	Parameter	Mean	Sd	HPD region	Mean	Sd	HPD region	Mean	Sd	HPD region
2	SMOKE	0.37	0.83	(-1.35 ,2.05)	0.23	0.66	(-1.10 ,1.54)	0.38	0.85	(-1.30,2.17)
	LWT	-0.52	1.61	(-3.76 ,2.52)	-0.55	1.59	(-3.73 ,2.41)	-0.55	1.62	(-3.65,2.79)
	UI	1.81	0.83	(0.18 ,3.51)	1.78	0.83	(0.30 ,3.59)	1.81	0.87	(0.27,3.72)

Table 5: Analysis of low birth weight data with two disease states, using full dataset. BSP stands for Bayesian semiparametric method whereas PBC and PBV stand for parametric Bayes methods assuming constant and varying stratum effects respectively.

		BSP			PB			PBV		
Logit	Parameter	Mean	Sd	HPD region	Mean	Sd	HPD region	Mean	Sd	HPD region
1	SMOKE	0.86	0.88	(-0.88,2.55)	0.55	0.78	(-0.84,2.18)	0.56	0.86	(-0.71,2.49)
	LWT	-0.92	1.31	(-3.63,1.51)	-1.03	1.34	(-3.50,1.83)	- 1.01	1.33	(-3.57,1.75)
	UI	0.19	0.69	(-1.11,1.49)	0.21	0.67	(- 1.10,1.55)	0.20	0.69	(-1.31,1.48)
Logit	Parameter	Mean	Sd	HPD region	Mean	Sd	HPD region	Mean	Sd	HPD region
2	SMOKE	0.54	1.04	(-1.46, 2.07)	0.13	0.93	(-1.85 ,1.88)	0.12	0.93	(-1.10,1.54)
	LWT	-0.43	1.62	(-3.98,2.38)	-0.59	1.63	(-3.75,2.59)	- 0.54	1.54	(-3.62,2.45)
	UI	1.82	0.86	(0.34,3.65)	1.80	0.83	(0.24,3.46)	1.87	0.82	(0.43,3.63)

Table 6: Analysis of low birth weight data with two disease states after deleting 40% observations on smoking completely at random. BSP stands for Bayesian semiparametric method whereas PB and PBV stand for parametric Bayes methods assuming constant and varying stratum effects respectively

References

Altham, P. M. E. (1971). The analysis of matched proportions. *Biometrika* **58**, 561–576.

Anderson, J. A. (1972). Separate sample logistic discrimination. *Biometrika* **59**, 19–35

Armstrong, B. G., Whittemore, A. S. and Howe, G. R. (1989). Analysis of case-control data with covariate measurement error: application to diet and colon cancer. *Statistics in Medicine* **8**, 1151–1163.

Armstrong, B. G. (2003). Fixed factors that modify the effects of time-varying factors: Applying the case-only approach. *Epidemiology* **14**, 467–472.

Ashby, D., Hutton, J. L. and McGee, M. A. (1993). Simple Bayesian analyses for case-controlled studies in cancer epidemiology. *Statistician* **42**, 385–389.

Breslow, N. E. (1996). Statistics in epidemiology: The case-control study. *Journal of the American Statistical Association* **91**, 14–28.

Breslow, N. E. and Cain (1988). Logistic regression for two-stage case-control data. *Biometrika* **75**, 11-20.

Breslow, N. E., Day, N. E., Halvorsen, K. T., Prentice, R. L., and Sabai, C. (1978). Estimation of multiple relative risk functions in matched case-control studies. *American Journal of Epidemiology* **108** , 299–307

Breslow, N. E. and Day, N. E. (1980). *Statistical Methods in Cancer Research*, Volume 1. Lyon, International Agency for Research on Cancer.

Carroll, R. J., Gail, M. H. and Lubin, J. H. (1993). Case-control studies with errors in covariates. *Journal of the American Statistical Association* **88**, 185–199.

Carroll, R. J., Wang, S. and Wang, C. Y. (1995). Prospective analysis of logistic case-control studies. *Journal of the American Statistical Association* **90**, 157–169.

Cheng, K. F. and Chen, J. H. (2005). Bayesian models for population based case-control studies when the population is in Hardy-Weinberg equilibrium. *Genetic Epidemiology* **28**, 183–192.

Cornfield, J. (1951). A method of estimating comparative rates from clinical data: applications to cancer of the lung, breast, and cervix. *Journal of the National Cancer Institute* **11**, 1269–1275.

Cornfield, J., Gordon, T. and Smith, W. W. (1961). Quantal response curves for experimentally uncontrolled variables, *Bulletin of the International Statistical Institute* **38**, 97–115.

Cox, D. R. (1966). A simple example of a comparison involving quantal data. *Biometrika* **53**, 215–220.

Day, N. E., and Kerridge, D. F. (1967). A general maximum likelihood discriminant. *Biometrics* **23**, 313–323.

De Santis, F. Perone Pacifico, M. and Sambcini, V. (2001). Optimal predictive sample size for case-control studies. Rapporto *Tecnico n. 17, Dipartmento di Statistica, Probabilita e Statistiche Applicate, Universita di Roma.*

Diggle, P. J., Morris, S. E. and Wakefield, J. C. (2000). Point-source modeling using matched case-control data. *Biostatistics* **1**, 89–105.

Escober, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* **90**, 577–588.

Gail, M. H., Pee, D., and Carroll, R. J. (2001). Effects of violations of assumptions on likelihood methods for estimating the penetrance of an autosomal dominant mutation from kin-cohort studies. *Journal of Statistical Planning and Inference* **96**, 167–177.

Ghosh, M. and Chen, M-H. (2002). Bayesian inference for matched case-control studies. *Sankhyā, B*, **64**, 107-127.

Godambe, V. P. (1976). Conditional likelihood and unconditional optimum estimating equations.*Biometrika* **63**,277–284.

Gustafson, P., Le, N. D., and Vallee, M. (2002). A Bayesian approach to case-control studies with errors in covariables. *Biostatistics* **3**, 229–243.

Hosmer, D. A. and Lemeshow, S. (2000). *Applied Logistic Regression.* John Wiley, second edition.

Ibrahim J. H. and Chen, M. H. (2000). Power prior distributions for regression models. *Statistical Science* **15**, 46–60.

Kaldor, J., Day, N., Band, P., Choi, N., Clarke, E., Coleman, M., Hakama M., Koch M., Langmark F., Neal, F., Petterson, F., Pompe-Kirin, V., Prior, P. and Storm, H. (1987) Second malignancies following testicular cancer, ovarian cancer and Hodgkin's disease: an international collaborative study among cancer registries, *International Journal of Cancer* **39**, 571–585.

Kim, I., Cohen, N. D. and Carroll, R. J. (2002). A method for graphical representation of effect heterogeneity by a matched covariate in matched case-control studies exemplified using data from a study of colic in horses. *American Journal of Epidemiology* **156**, 463–470.

Khoury, M. J., Beaty, T. H., and Cohen, B. H. (1993). *Fundamentals of genetic epidemiology.* Oxford University Press (Oxford).

Lindley, D. V. (1964). The Bayesian analysis of contingency tables. *The Annals of Mathematical Statistics* **35**, 1622–1643.

Lipsitz, S. R., Parzen, M. and Ewell, M. (1998). Inference using Conditional Logistic Regression with Missing Covariates. *Biometrics* **54**, 295–303.

Little, R. J. A., and Rubin, D. B. (1987). *Statistical analysis with missing data.* New York: John Wiley.

Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute* **22**, 719–748.

Marshall, R. J. (1988). Bayesian analysis of case-control studies. *Statistics in Medicine* **7**, 1223–1230.

McShane, L. M., Midthune, D. N., Dorgan, J. F., Freedman, L. S. and Carroll, R. J. (2001). Covariate measurement error adjustment for matched case-control studies. *Biometrics* **57**, 62-73.

Müller, P., Parmigiani, G., Schildkraut, J. and Tardella, L. (1999). A Bayesian hierarchical approach for combining case-control and prospective studies. *Biometrics* **55**, 858–866.

Müller, P. and Roeder, K. (1997). A Bayesian semiparametric model for case-control studies with errors in variables. *Biometrika* **84**, 523–537.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9, 249–265.

Nurminen, M. and Mutanen, P. (1987). Exact Bayesian analysis of two proportions. *Scandinavian Journal of Statistics* **14**, 67–77.

Oakes, D. (1986). Semiparametric inference in a model for association in bivariate survival data. *Biometrika* **73**, 353–361

Oakes, D. (1989). Bivariate survival models induced by frailties. *Journal of the American Statistical Association* **84**, 487–493.

Paik, M. C. and Sacco, R. (2000). Matched case-control data analyses with missing covariates. *Applied Statistics* **49**, 145–156.

Prentice, R. L. and Breslow, N. E. (1978). Retrospective Studies and Failure Time Model.

Biometrika **65**, 153–158.

Prentice, R. L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika* **66**, 403–411.

Prentice, R. L. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika* **73**, 1–11.

Raftery, A. E., and Richardson, S. (1996). Model selection for generalized linear models via GLIB: Application to nutrition and breast cancer. *Bayesian Biostatistics*, 321–353.

Rathouz, P. J., Satten, G. A. and Carroll, R. J. (2002). Semiparametric inference in matched case-control studies with missing covariate data. *Biometrika* **89**, 905–916.

Rice, M. K. (2003). Full-likelihood approaches to misclassification of a binary exposure in matched case-control studies. *Statistics in Medicine*, **22**, 3177-3194.

Rice, M. K. (2004). Equivalence between conditional and mixture approaches to the rasch model and matched case-control studies, with application. *Journal of the American Statistical Association* **99**, 510–522.

Roeder, K., Carroll, R. J., and Lindsay, B. G. (1996). A semiparametric mixture approach to case-control studies with errors in covariables. *Journal of the American Statistical Association* **91**, 722–732.

Rubin, D. B. (1981). The Bayesian bootstrap. *The Annals of Statistics* **9**, 130–134.

Satten, G. A. and Carroll, R. J. (2000). Conditional and unconditional categorical regression models with missing covariates. *Biometrics* **56**, 384–388.

Satten, G. A. and Kupper, L. (1993). Inferences about exposure-disease association using Probabiliy-of-Exposure Information. *Journal of the American Statistical Association* **88**, 200–208.

Satten, G. A., Flanders, W. D. and Yang, Q. (2001). Accounting for Unmeasured Population Substructure in Case-Control Studies of Genetic Association Using a Novel Latent-Class Model. *Ameriacan Journal of Human Genetics* **68**, 466–477.

Scott, A. J. and Wild, C. J. (1986). Fitting logistic models under case-control or choice-based sampling. *Journal of the Royal Statistical Society, B*, **48**, 170-182.

Seaman, S. R. and Richardson, S. (2004). Equivalence of prospective and retrospective models in the Bayesian analysis of case-control studies. *Biometrika* **91**, 15–25.

Seaman, S. R. and Richardson, S. (2001). Bayesian analysis of case-control studies with categorical covariates. *Biometrika* **88**, 1073–1088.

Shih, J. H. and Chatterjee, N. (2002). Analysis of survival data from case-control family studies. *Biometrics* **58**, 502–509.

Sinha, S., Mukherjee, B. and Ghosh, M. (2005b). Modeling association among multivariate exposures in matched case-control study. *Preprint.*

Sinha, S., Mukherjee, B. and Ghosh, M., Mallick, B. K. and Carroll, R. (2005a). Semiparametric Bayesian modeling of matched case-control studies with with missing exposure. *Journal of the American Statistical Association*, to appear.

Sinha, S., Mukherjee, B. and Ghosh, M. (2004). Bayesian semiparametric modeling for matched case-control studies with multiple disease states. *Biometrics* **60**, 41–49.

Wacholder, S. (1996). The case-control study as data missing by design: estimating risk difference. *Epidemiology* **7**, 144–150.

Walsh R. D. (1994). Effects of maternal smoking on adverse pregnancy outcomes: Examination of the criteria of causation. *Human Biology* **66**, 1059–1092.

West, M., Müller, P., and Escobar, M. D. (1994). Hierarchical priors and mixture models, with application in regression and density estimation. *Aspects of Uncertainty. A Tribute to D. V. Lindley*, 363–386.

Zelen, M. and Parker, R. A. (1986). Case-control studies and Bayesian inference. *Statistics in Medicine* **5**, 261–269.